

Modeling Attitude Change from Political Dialogues

Peter Duggins

Center for Advanced Modeling in the Social, Behavioral, and Health Sciences, Johns Hopkins University

July 1, 2014

I use an agent-based simulation to model how individual and societal attitude change arise from the mutual interactions that occur within political dialogues. Agents converse with members of their social network, express their political attitudes, and update their beliefs in a manner that reflects their personal convictions, their tolerance for dissimilarity, and local social norms. The model extends previous work by endowing agents with multiple, theoretically motivated heuristics for reevaluating and expressing their opinions, including homophily, attitude strength, and conformity. I demonstrate that novel macroscopic phenomenon emerge from the interactions of these cognitively and socially refined agents. Specifically, I find that (a) interactions between persuadable and closed-minded agents drive non-monotonic changes in the societal distribution of opinions; (b) extremists instigate neighborhood polarization that propagates outward through undecided agents; (c) preference falsification produces non-representative public norms that obscure attitude diversity and consequently alter dynamics of attitude change. I conclude by discussing the political implications of the results, suggesting an experiment to validate the model's findings, and proposing extensions for future work.

Keywords: Agent-Based Model, Opinion Dynamics, Attitude Change, Social Influence, Homophily, Conformity

INTRODUCTION

Attitudes, defined as evaluative judgments on subjective issues, lie at the heart of political thinking and behavior. Attitudes are simultaneously persistent and mutable: stable evaluations are stored in memory via associative links, but context alters how these attitudes are constructed and communicated to others (Bohner and Dickel 2011). The formation, reevaluation, and expression of attitudes are fundamental in many political domains. For example, when large groups of people reconsider their attitudes about issues like gay marriage and legalized marijuana, public opinion shifts and grassroots referenda may alter state policy. Similarly, when influential individuals like congressmen or supreme court justices adopt a new stance on a controversial issue, they may implement top-down changes in policy by enacting legislation or setting legal precedents. Previous work (Hegselmann and Krause 2002; Deffuant et al. 2002; Dandekar et al. 2013) has established the importance of factors like social influence, tolerance of dissimilarity, and personal convictions in promoting attitude change, but new models are needed to synthesize these factors into a coherent, quantitative account of how changing political beliefs give rise to collective political phenomena.

In this paper, I propose an agent-based model of attitude change that extends previous work by endowing agents with multiple, theoretically motivated heuristics for updating and expressing their attitudes. These heuristics account for the similarity between agents political beliefs when calculating social influence (homophily), for the personal convictions held by agents when determining persuadability (attitude strength), and for the social norms of the group when expressing attitudes publicly (conformity). Agents interact with members of their social networks through political dialogues, exchange opinions about a political issue, and undergo attitude change. I demonstrate that the cognitive and social depth afforded by these heuristics produces three novel emergent behaviors. First, the distribution of political attitudes within the population changes non-monotonically over time. Second, neighborhoods develop strong political norms and grow in size as moderates on the periphery influence nearby persuadable agents. Third, preference falsification conceals the dissent of individuals who privately reject political norms. These results indicate that, in order to explain the complexities of societal attitude change, we must understand and model how individuals change their opinions based on internal characteristics, external stimuli, social context, and social networks.

Theories of Attitude Change

In general, a robust theory of macroscopic attitude change must meet the following criteria. First, it must use social psychology and cognitive science to build a mechanistic model of attitude change at the individual level. Second, it must describe how people interact socially by specifying their social networks and the dynamics of political interactions. Third, it must show how individual attitude change results from repeated interpersonal interactions.

Fourth, it must generate the macroscopic phenomenon of interest from accumulated attitude change within a population. Fifth, it must be sufficiently quantitative to make falsifiable predictions and transparent enough to allow replication and meaningful interpretation.

Many current theories fail to meet one or more of these criteria. Leadership theories, which emphasize the role of particular people, groups, or institutions in societal movements (Morris and Staggenborg 2004), often lack necessary the broadly-applicable framework, falsifiability, and predictive capability. Psychosocial theories, which empirically show that individuals are simultaneously motivated to form accurate attitudes, socially acceptable attitudes, and self-consistent attitudes (Wood 2000; Cialdini and Goldstein 2004), rarely quantify the attitude change that results when two or more experiences exert opposing influence. Opinion dynamics models, which mathematically describe attitude change using statistical mechanics or agent-based formalisms, are quantitative and reproducible, but have yet to incorporate sufficient cognitive and social depth (see below).

Studying Attitude Change

To understand the origins of attitude change, and study its consequences, we must first answer the question, “What prompts individuals to change their attitudes on political issues?” Broadly speaking, people are exposed to novel political information and arguments from valued reference groups, such as members of their social networks or mass media, that spur them to reevaluate their beliefs. An individual’s resulting attitude change depends in some way on his internal attributes, such the strength of his political beliefs and his motivation to form accurate attitudes, on external attributes of the stimulus and its source, such as informational content and trustworthiness, and on context, such popular approval and pressure to conform.

In general, people account for both evidence and subjective evaluations when forming and updating their attitudes. However, in the domain of politics, the existence of “objective information is tenuous, given that most facts can be interpreted differently by two ideologically disparate individuals: a statistic stating the incidence of veiling by Muslim women may be perceived by a feminist as a sign of gender inequality and by a Islamic priest as an indication of modesty. Furthermore, direct personal experience with political objects is rare: few people directly perceive the advantages and disadvantages of a states foreign policy. For these reasons, many people base their political attitudes on subjective evaluations made by valued individuals (Wood 2000; Kuran 1997) rather than on evidence they gather themselves. This study models how individuals incorporate the subjective statements made by others into their political worldview.

Dialogues are one specific instance of political interaction in which people connect to each other through social networks, engage in conversations where they express their beliefs and consider the evaluative statements of others, and update their beliefs based on what they hear. When a person engages in such a dialogue, his attitude change is proportional to the total weighted *influence* exerted on him, which depends on the number of attitude statements, the

persuasiveness of each statement, and his political receptiveness (Latane 1981; DeGroot 1974). Although dialogues are a relatively specific domain, this model framework can be applied to other sources of political influence, such as the media, or to other forms of political interaction, such as diplomatic negotiations between states.

Cognitive Heuristics

The extent to which an individual is persuaded by others subjective evaluations depends on numerous interdependent factors. These include: previous beliefs and dissonance reduction (Petty et al. 1997); motivations to be accurate, self-consistent, and socially accepted (Wood 2000; Cialdini and Goldstein 2004); issue framing, emotional arousal, and cognitive elaboration (Bohner and Dickel 2011; Gawronski and Bodenhausen 2006); self-esteem (Pool et al. 1998); social norms (Betz et al. 1996); age Visser and Krosnick (1998); and more. Social psychology has shown that individuals often use simplifying heuristics to make rapid judgments about the validity of others political statements. *Homophily* describes individuals attraction to similarity: people who hold comparable opinions are more likely to form friendships and be persuaded by each others arguments (McPherson et al. 2001). *Attitude strength* highlights individuals reliance on entrenched beliefs: people who have deliberated and formed strong opinions on a political issue are less likely to change their mind than uninformed, undecided, or neutral individuals (Ajzen 2001; Taber and Lodge 2006). *Conformity* outlines individuals desire to gain social approval and maintain relationships: people who hold opinions that run contrary to public social norms will hide or even misrepresent their true beliefs in order to avoid ostracism (Wood 2000; Cialdini and Goldstein 2004; Kuran 1989). Empirical studies have shown that when individuals apply these mental shortcuts to the evaluation of political arguments, their *biased assimilation* of information may lead them to make complex or “irrational belief changes. For example, individuals may adopt more extreme attitudes when exposed to a balanced or counter-attitudinal set of arguments, a phenomenon referred to as *attitude polarization* (Lord et al. 1979; Miller et al. 1993; Taber and Lodge 2006).

Social Networks

A robust theory of attitude change must account for *social networks* in order to simulate the occurrence and frequency of interactions between individuals in a population. One method for generating plausible social networks places agents on a two-dimensional grid and lets each one form network connections with other agents who lie within their “social reach. Social reach is determined by the radial parameter R , a Euclidian distance representing the geographical constraints on an agents social network, and by extension the connectivity of the population as a whole. Social reach models grow networks with several features that classical network models (perfect mixing, Moore neighborhoods, and preferential attachment networks) fail to achieve: these networks are size limited, spatially constrained, heterogeneous in size and density, have a low whole network density, positive degree assortativity, and short path lengths (Hamill and Gilbert 2010). One extension of the social reach framework simulates gatherings in

which individuals invite members of their social reach networks to spaces where community interactions take place and new social linkages are made, enabling the formation of social triangles and other network features absent in the original model (zu Erbach-Schoenberg et al. 2013).

Opinion Dynamics

Opinion dynamics models are agent-based models that investigate attitude change within an artificial population of interacting individuals following simple opinion-updating rules. One commonly employed assumption in these models is *bounded confidence* (BC), a mathematical realization of homophily which claims that individuals whose opinion differ by more than a certain threshold are too dissimilar to persuade one another. BC models (Hegselmann and Krause 2002; Deffuant et al. 2002; Dandekar et al. 2013; Salzarulo 2006; Jager and Amblard 2005) have demonstrated that agents tolerance for dissimilar opinions, represented as the confidence threshold, affects whether society comes to agreement on a contentious issue (convergence) or splits into two or more opinions (divergence). Although BC is a significant improvement from qualitative descriptions of homophily, in that it mathematically explicates a relationship between attitude change and social influence, it implies that persuasiveness is “all or nothing” and eliminates the richness of interaction possible with homophily. Furthermore, BC models fail to account for other attitude change heuristics, including attitude strength and conformity, nor do they typically employ plausible social networking between agents. Most importantly, the behavior of BC agents is entirely characterized by monotonic convergence to one or more opinion attractors; they do not produce either non-monotonic dynamics of attitude change or the persistence of (a continuum of) heterogeneous opinions, two macroscopic results which are readily observable in populations of real individuals who exchange political beliefs. Opinion dynamics models have only begun to address the immense complexity of individuals attitude change and the resulting collective political phenomena.

MODEL SPECIFICATION

Overview

I propose a psychologically motivated, socially constrained, agent based computational model of attitude change. Agents begin the simulation with an opinion on a subjective political issue and a social network filled with nearby agents. Agents converse with each other in dialogues in which they express their attitude and listen to the opinions stated by others. After the dialogue, agents weigh the opinions they heard and finally reevaluate their own attitudes. In a series of computational experiments, I manipulate the parameters and equations governing how agents weigh each others opinions and how they choose to express their own attitudes. Each manipulation reflects a unique

assumption about the factors influencing attitude change, including internal attributes like open-mindedness, external attributes like persuasiveness, and contextual factors like social norms. For each experiment, I theoretically motivate the manipulation, examine and explain the resulting model behavior, and discuss the social and political implications of the result.

Skeletal Version

Agents hold a *private belief* represented as a continuous value on a one-dimensional political spectrum, $P_i(t) \in (-1, 1)$, where i is the agent index. Initial opinions are drawn from a uniform distribution. Agents sparsely populate a two-dimensional grid of size $D \times D$, and form network *links* with other agents within $distance_{ij} < R$. Each time step, one agent initiates a political *dialogue* with members of his social network. During a dialogue, each participant expresses his private opinion $P_i(t)$ once in random succession. When the dialogue ends, each participant updates $P_i(t)$ based on the dialogue's *impact* $I_i(t)$:

$$P_i(t) = P_i(t-1) + I_i(t). \quad (1)$$

Impact is proportional to the weighted mean of participants expressed attitudes, which in the skeletal version are simply their private opinions $P_j(t)$:

$$I_i(t) = \frac{1}{N} \sum_j^N w_{ij} (P_j(t) - P_i(t)) \quad (2)$$

where N is the number of dialogue participants (equal to the number of agents in agent.i's network), $P_j(t)$ is the opinion that agent.j expresses in the dialogue, and w_{ij} is the weight that agent.i assigns to agent.j's statement. Equation 1 and 2 imply that agent.i shifts his political attitude in the direction of each $P_j(t)$ by an amount proportional to the statement's weight, which may reflect its persuasiveness (Equation 5), the receiving agents receptiveness (Equation 7), or the social context (Equation 10). The total impact of the dialogue on agent.i is the *mean directed shift* of his private opinion $P_i(t)$.

Metrics

In order to study attitude change at the societal scale, we wish to answer the question, "Given some initial distribution of opinions within the population, how does this distribution change over time?" I track the distribution of private opinions using histograms: I partition the total range of opinions into 200 bins of width $\delta P = 0.05$, then count the number of agents whose opinion falls within each bin. The resulting metric is recorded over the course of time, permitting an examination of the speed and extent of societal attitude change.

To complement this metric, I divide the spectrum of political attitudes into three categories of ideological

strength:

$$\begin{aligned}
\text{Centrists} &= \text{agents with } 0 < |P_i(t)| \leq 0.33 \\
\text{Moderates} &= \text{agents with } 0.33 < |P_i(t)| \leq 0.66 \\
\text{Extremists} &= \text{agents with } 0.66 < |P_i(t)| \leq 1.
\end{aligned} \tag{3}$$

To quantify the extent of neighborhood polarization (spatial clustering of opinions), I complement the generated patterns of opinions with the *Moran's I* statistic, which measures spatial autocorrelation of opinions (Moran 1950). Moran's I is defined as:

$$M_I = \frac{N}{\sum_i \sum_j L_{ij}} \frac{\sum_i \sum_j L_{ij} * (P_i(t) - \overline{P(t)}) (P_j(t) - \overline{P(t)})}{\sum_i (P_i(t) - \overline{P(t)})^2} \tag{4}$$

where N is the population size and L_{ij} is an element of a matrix of spatial links, such that $L_{ij} = 1$ if agent j is in agent i 's network, and is zero otherwise. Moran's I ranges from 0, indicating random dispersion of opinions, to 1, indicating perfect correlation of opinions.

The simulation begins by initializing 3000 citizens with randomized opinions $P_i(0) \sim U(-1, 1)$, randomized spatial positions on a 200×200 grid, and networks of social reach $R = 10$. Agents converse in dialogues for a fixed number of time steps with model metrics assessed every 250 steps. Appendix B lists the parameter values for each experiment. Unless otherwise reported, all model metrics (with the exception of the screenshots) are averaged over 100 stochastic realizations to ensure generality. The simulation was implemented in Netlogo 5.0.4 (Wilensky 1999).

RESULTS

Homophily Drives Convergence and Divergence

I begin by assuming that, when an agent weighs an opinion he hears in a dialogue, he considers the content of the statement in relation to his own attitude and assigns it some "persuasiveness. Specifically, agents use a homophily-based heuristic to calculate weight w_{ij} :

$$w_{ij} = 1 - s_h * \Delta P_{ij}(t) \tag{5}$$

$$\Delta P_{ij}(t) = |P_j(t) - P_i(t)| \tag{6}$$

where s_h , the salience of homophily, is a parameter governing agents intolerance of dissimilar opinions. Equation 5 implies that the weight of agent j 's statement decreases as the difference between agent i 's private opinion and

agent_js expressed statement grows¹: the more dissimilar the two agents opinions, and the higher their intolerance, the less attitude change will result from their dialogue. Equation 5 is a departure from BC models in that weight is a *continuous* rather than threshold function of $\Delta P_{ij}(t)$, allowing a greater diversity of interactions between agents. In this experiment, I show how tolerance of political diversity leads to societal consensus on a neutral opinion, while intolerance of diversity polarizes society into two moderate parties.

In order to assure that weights are non-negative, I artificially truncate w_{ij} to the interval $(0, 1)$. One consequence of this truncation is that the parameter s_h determines not only the magnitude of attitude change between two agents whose opinions differ by ΔP_{ij} , but also the *threshold* beyond which attitude differences become so great that *persuasion cannot occur*. Rearranging Equation 5 and setting $w_{ij} = 0$, persuasion disappears when agents attitudes differ by more than $\Delta P_{ij} > \frac{1}{s_h}$.

When agents intolerance for dissimilar opinions is low, $s_h \leq 1$, agents who converse in dialogues will find each others statements persuasive, $w_{ij} > 0$. Mutual agreement between ideologically distant individuals will push agents to adopt private opinions closer to the groups mean opinion. Because each agent belongs to many social networks, because social networks are well connected within the population, and because the mean initial opinion of the population is zero, society will converge to a single central opinion $P \rightarrow 0$, as shown in Figure 1a. On the other hand, when agents intolerance is high, $s_h = 2.5$, agents will not be persuaded by dissimilar opinion statements ($w_{ij} = 0$ when $\Delta P_{ij} > \frac{1}{s_h}$), and will only shift their opinions towards neighbors who share similar beliefs. Under these conditions, the distribution of private opinions will bifurcate into two *parties* ($P \rightarrow -0.5$ and 0.5), groups of agents who converse with one another but who cannot persuade members of the other group, Figure 1b². These results are consistent with findings from other opinion dynamics models, which use homophilous bounded-confidence heuristics to show that agents tolerances determine whether opinions converge or diverge (Deffuant et al. 2000; Hegselmann and Krause 2002; Jager and Amblard 2005).

Although two distinct parties emerge when intolerance is high, opinions remain well-mixed within the population, as shown in Figure 2. Mixing arises because polarization occurs within networks rather than between neighborhoods: each network diverges from initial uniformity until it contains members of two parties that cease to influence one another³. While the bifurcation of an intolerant society into two opposing parties is consistent with intuition, one might expect to observe a corresponding spatial segregation between these parties along neighborhood lines. This begs the question “what drives neighborhood polarization?”

¹An alternative hypothesis supposes just the opposite: that agents are most persuaded by arguments that are radically different from their own because they are surprising and highly salient. This suggests a mechanism for jolting agents out of complacent beliefs and is undoubtedly responsible for the occasional “political revelation”. Although the seeds of doubt planted by radical arguments may lead to significant attitude change in some circumstances, these attitudes are (I suspect) statistically more likely to be disregarded as “crazy” than highly persuasive. For tractability I will assume that persuasiveness declines monotonically with opinion dissimilarity according to Equation 5

²Other values of tolerance produce three or more parties ($s_h = 4$) or dynamics where two parties initially form but a centrist party emerges and eventually becomes universally accepted ($s_h = 2$).

³Under these conditions, space and network are irrelevant; dispensing with localized interaction and assuming perfect mixing within the population results in the same opinion distribution.

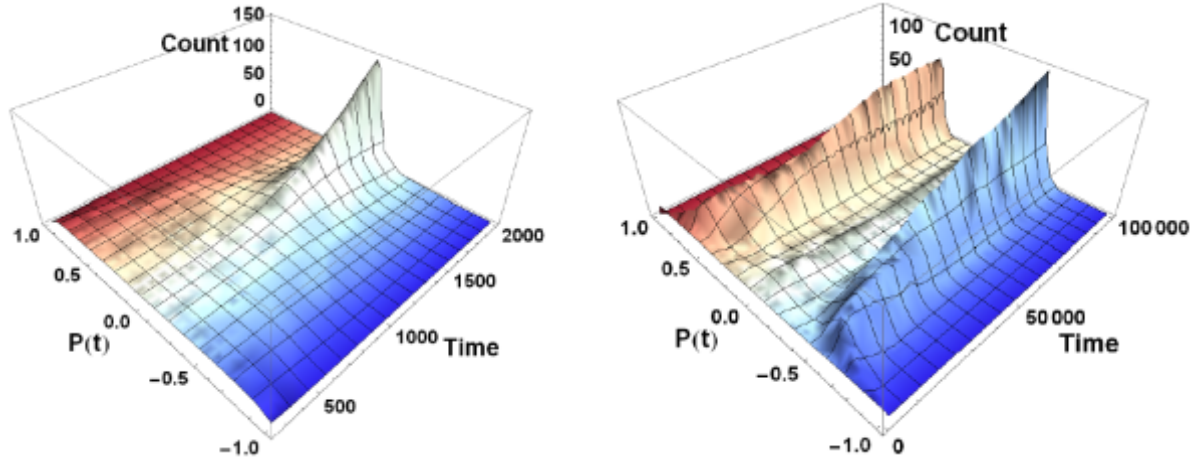


Figure 1: When intolerance of dissimilar opinions is low, agents are persuaded by arguments across a large ideological spectrum, and mutual agreement causes society to converge to a centrist opinion (left); when intolerance is higher, agents disregard dissimilar statements, and society bifurcates into two moderate parties (right). Given enough time, all opinions converge *absolutely* to these attractor(s), leaving no diversity in political attitudes.

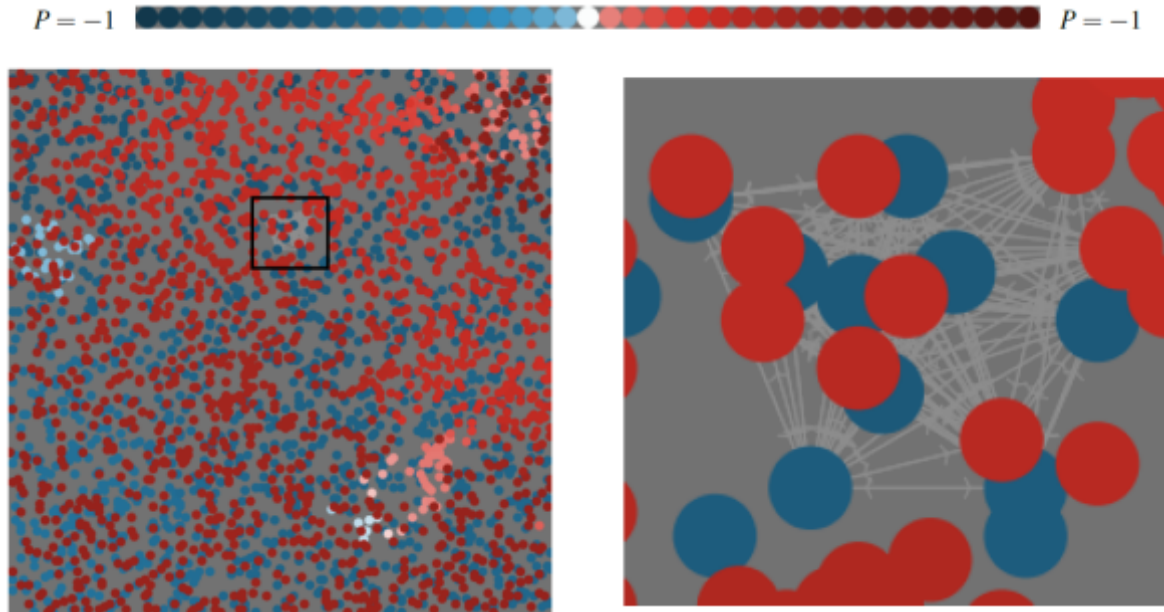


Figure 2: Homophily and high intolerance drive the in-group convergence of red ($P = 0.5$) and blue ($P = -0.5$) agents inside individual networks. This causes spatial mixing within the population ($M_I = 0.005 \pm 0.01$ at $t = 50,000$) despite the centralization of the two moderate parties shown in Figure 1b.

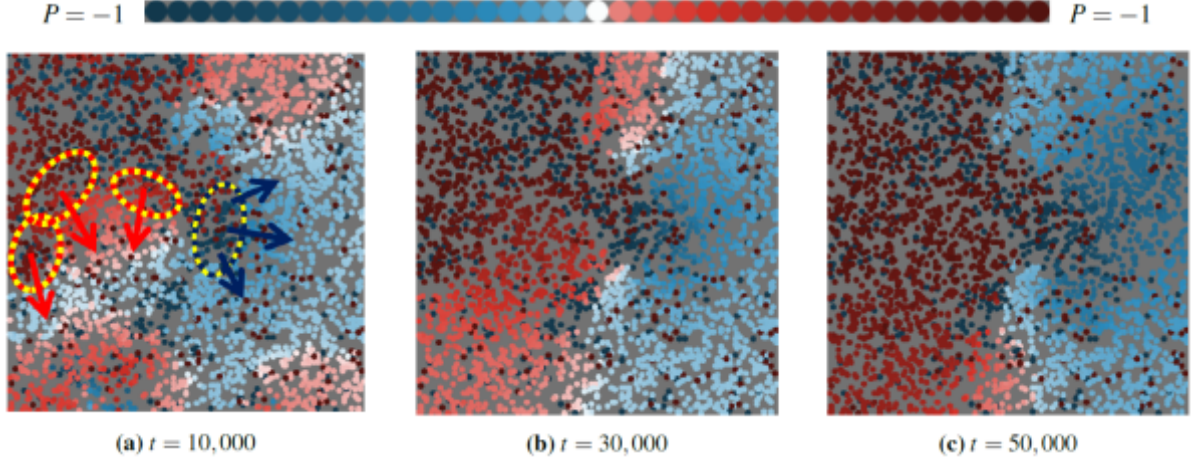


Figure 3: An epidemic of radicalization spreads spatially outward from extremist neighborhoods (circled in (a)) through neutral agents, causing neighborhood polarization. Morans I increases as polarization spreads: $M_I = 0.08, 0.24, 0.42$.

Extremists Polarize Neighborhoods

In this experiment, I discuss why attitude rejection fosters extremism and demonstrate how extremists drive neighborhood polarization. Rejection occurs when an individual is so appalled by an opinion that his attitude shifts *away from* it. I incorporate rejection by allowing “rejector” agents to assign *negative* weight to other’ statements, $w_{ij} = (-1, 1)$. When rejectors are exposed to a diverse set of expressed beliefs, they shun those which are most dissimilar from their own and quickly adopt extremist attitudes (see Appendix A for details). In essence, rejectors undergo *attitude polarization from biased assimilation* where normal agents would not ⁴.

Introducing rejectors into the population significantly alters the dynamics of societal attitude change. In a tolerant society ($s_h = 1.25$) where $N_r = 15\%$ of agents are rejectors, most agents, regardless of ideological strength (defined by Equation 3), still come to a consensus around the politically neutral centrism, $P \rightarrow 0$. However, rejector agents undergo attitude polarization when exposed to dissimilar opinions and consequently radicalize, $|P| \rightarrow 1$. By chance, some neighborhoods will contain disproportionate numbers of (now extremist) rejectors. Although radical neighborhoods have no chance of persuading other rejectors to change their attitudes, sustained influence from these neighborhoods will radicalize surrounding centrists. As neutral agents adopt moderate attitudes, they in turn polarize surrounding centrists. A resurgence of extremism propagates through ‘undecided agents, eventually overtaking society, Figure 3 and 4. Once this cascade has begun, it is the moderates that continue to drive divergence: artificially removing all extremists at $t = 20,000$ does not stop absolute societal polarization. In general, closed-minded extremists instigate neighborhood polarization by radicalizing persuadable agents, who then propagate strong opinions spatially outwards from the extremist core. Subsequent experiments show that this

⁴When two rejectors of opposing initial opinion are placed next to each other, they undergo positive feedback polarization, and become thoroughly entrenched in their beliefs.

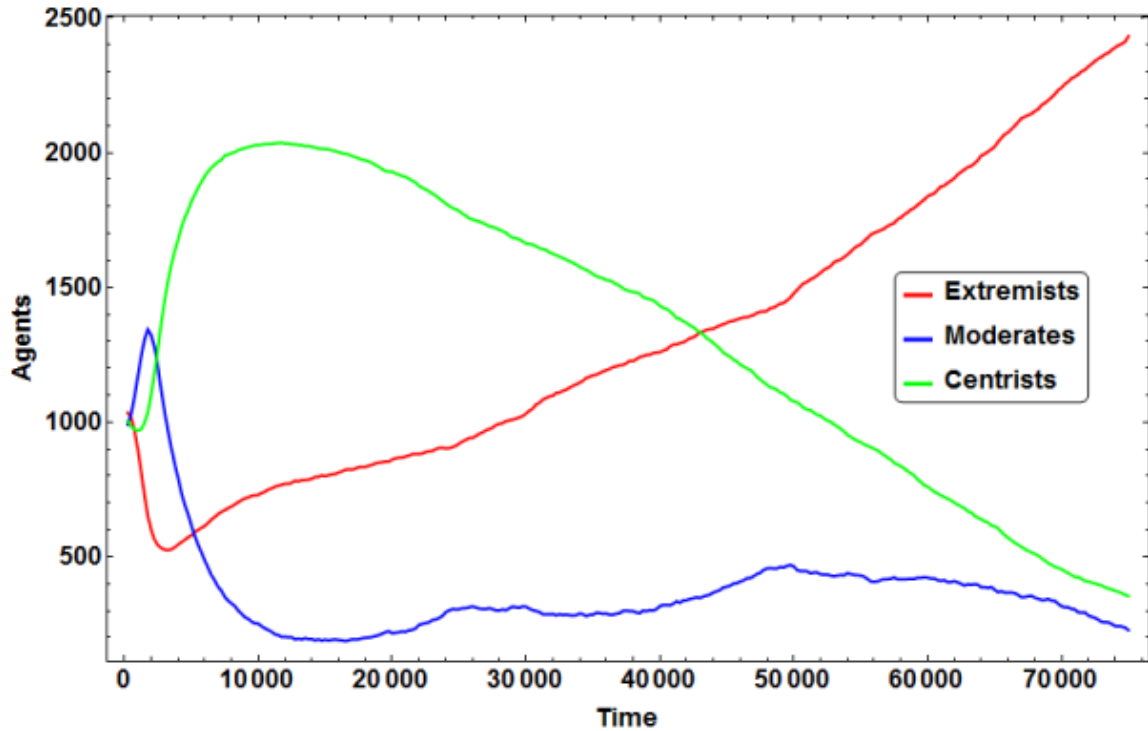


Figure 4: Initially, tolerance causes a strong societal trend towards centrism while rejectors turn to extremism. Later, sustained influence from closed-minded extremists radicalizes undecided agents, initiating a wave of polarization that spreads until all centrist and moderate agents have adopted extremism.

trend holds under several definitions of extremism and persuadability.

Attitude Strength Produces Diversity

In this experiment, I continue exploring the dynamics between extremists and susceptible agents by introducing *attitude strength*, the tendency for strongly-opinionated individuals to resist attitude change. Like homophily, attitude strength is a robust determinant of attitude change (Bohner and Dickel 2011; Gawronski and Bodenhausen 2006). In this case, an individual's degree of attitude change is determined by his personal attributes (closed-mindedness) rather than external attributes of the message or its source (homophily). Attitude strength is manifested by an individual's ability to quickly recall salient attitudes from memory, cognitively manipulate them to address novel counterarguments, and accurately express their point of view (Petty et al. 1997). Here, I show that when agents' closed-mindedness increases with the strength of their attitudes, a diversity of political beliefs can be sustained within society.

Despite its theoretical significance, opinion dynamics models have yet to incorporate attitude strength. I address this gap by altering how agents determine an argument's persuasiveness. The homophilous calculation of weight,

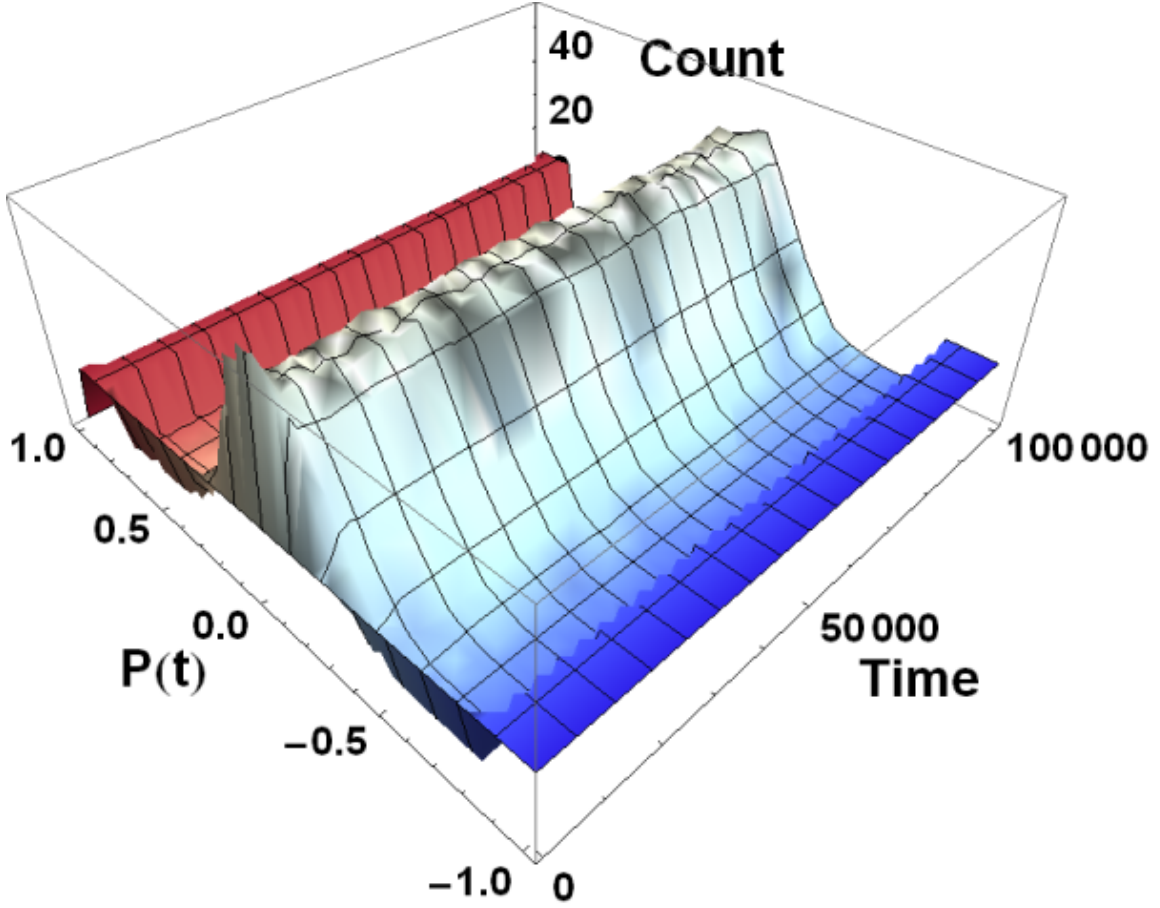


Figure 5: A diversity of opinions within society persists indefinitely when agents receptiveness to political influence declines with the strength of their existing beliefs. Individuals opinions fluctuate between centrism and extremism, but on average the distribution of private opinions remains normally distributed around $P = 0$ with two isolated, immovable extremist parties.

Equation 5, is replaced by an attitude-strength calculation that reflects agents political receptiveness:

$$w_{ij} = 1 - s_a * |P_i(t)| \quad (7)$$

where s_a , the salience of attitude strength, is a parameter governing agents' "closed-mindedness", with small s_a indicating high receptiveness to influence. Equation 7 implies that if agent i holds an extreme private belief, he will resist influence from any argument he hears, similar or dissimilar, and maintain a constant opinion. Conversely, if agent i 's private beliefs are neutral, he is said to be "undecided" on the issue, and will be highly susceptible to radicalizing influence⁵(see Appendix A).

Agents behaving according to Equation 7 experience two competing forces: a pull towards centrism resulting

⁵Note that, in reality, having a neutral political opinion does not necessarily imply that an individual is uninformed/undecided about a political issue or that his opinion is weakly held. For example, an individual may be so disenfranchised about a political issue that he refuses to take sides as a matter of principle, or may deem the issue of such low importance that forming a concrete opinion is a waste of time. Nonetheless, I contend that on average, strongly opinionated individuals are less susceptible to influence than weakly opinionated individuals.

from dialogues which have neutral political norms; and a pull towards extremism from closed-minded radicals scattered across the geographic space. In the previous experiment, extremists maintained radical views through their rejection of moderating influence, but managed to induce extreme views in nearby centrists, thereby increasing the pool of extremists and instigating a cascade of polarization. In this experiment, extremists maintain radical views through closed-mindedness, but as they induce extreme views in nearby centrists, those centrists become increasingly closed-minded and resist further radicalization. Consequently, the extremist population does not grow at a sufficient rate to initiate a cascade of polarization, and centrist influence dominates. However, as moderates move back towards centrism, they become increasingly open-minded, and are once-again susceptible to extremist influence. The consequence is that agents attitudes fluctuate indefinitely between extremism and centrism, producing a *stable, diverse distribution of opinions within society*, Figure 5. Sustained diversity is a novel result in the context of BC models, in which opinions always converge absolutely to one or more attractors given sufficient time.

Homophily and Attitude Strength

In the hopes of capturing the dynamics of homophily and the stability of attitude strength, I now merge the two into a composite heuristic for attitude change. Many mathematical combinations of Equations 5 and 7 are possible and would likely yield interesting behavior; for tractability, I use a simple linear combination:

$$w_{ij} = 1 - s_h * \Delta P_{ij}(t) - s_a * |P_i(t)|. \quad (8)$$

I initialize a population with moderate intolerance and close-mindedness, $s_h = s_a = 0.75$. Extremists and centrists initially form distinct groups, as in Figure 5, but extremists now induce moderate views in neutral agents. A burgeoning moderate population recruits from both the extremist and centrist parties, but tolerant interactions between opposing moderate agents may lead to consensus and a return to centrism. If this occurs, the moderate population declines until extremist again polarize the centrist population. Competing influence between closed-minded extreme parties, populace moderate parties, and a susceptible centrists party inspire non-monotonic, erratic opinion dynamics that continue for hundreds of thousands of dialogues. This behavior demonstrates how increasing agents cognitive depth, by accounting for both homophily and attitude strength, can generate novel patterns of attitude change and multimodal/non-symmetric opinion distributions, Figure 6 and 7.

The combination of homophily and attitude strength also produces highly polarized neighborhoods, as shown in Figures 9 and 8. Within these neighborhoods, extreme agents are located in the center, isolated from dissimilar influence, and moderates are located nearer the border, where dialogues between opposing parties encourage the adoption of political neutrality. As competition occurs on the border, parties may gain or lose territory, causing

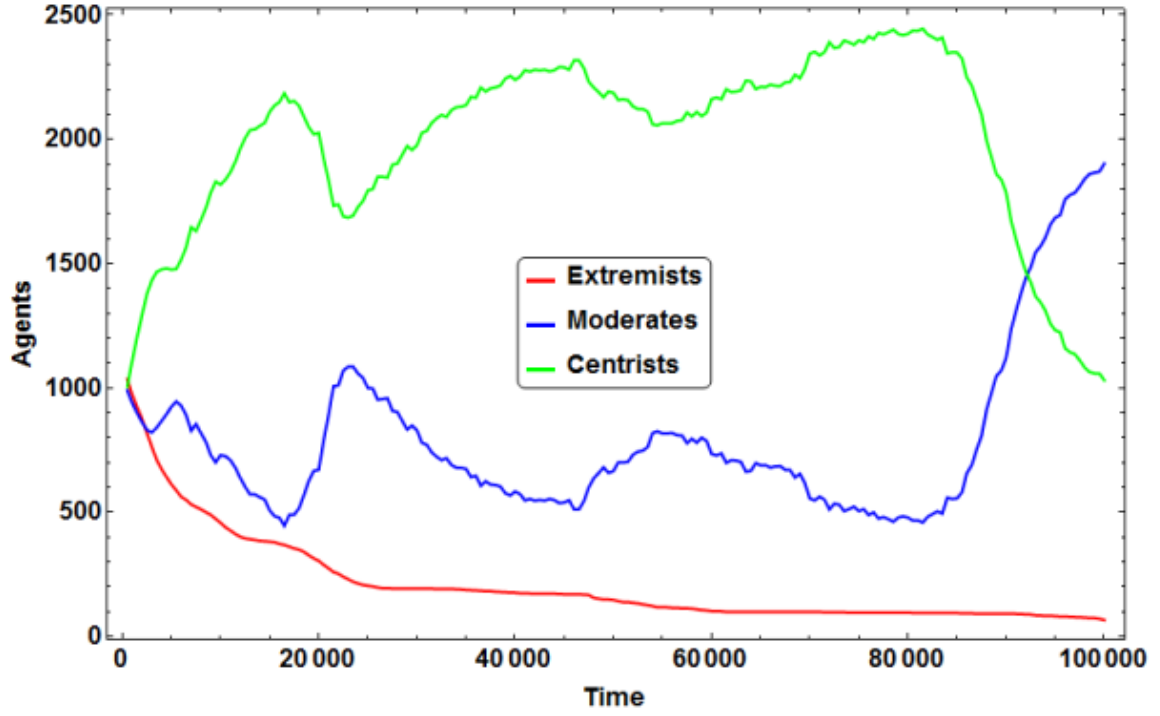


Figure 6: Non-monotonic dynamics between extremist, moderate, and centrist parties emerge when homophily and attitude strength jointly determine agents attitude change. In the reported realization, centrists dominate the political landscape despite surges in the moderate population at $t = 25,000$ and $t = 55,000$. At $t = 90,000$, moderates gain the upper hand and overtake society.

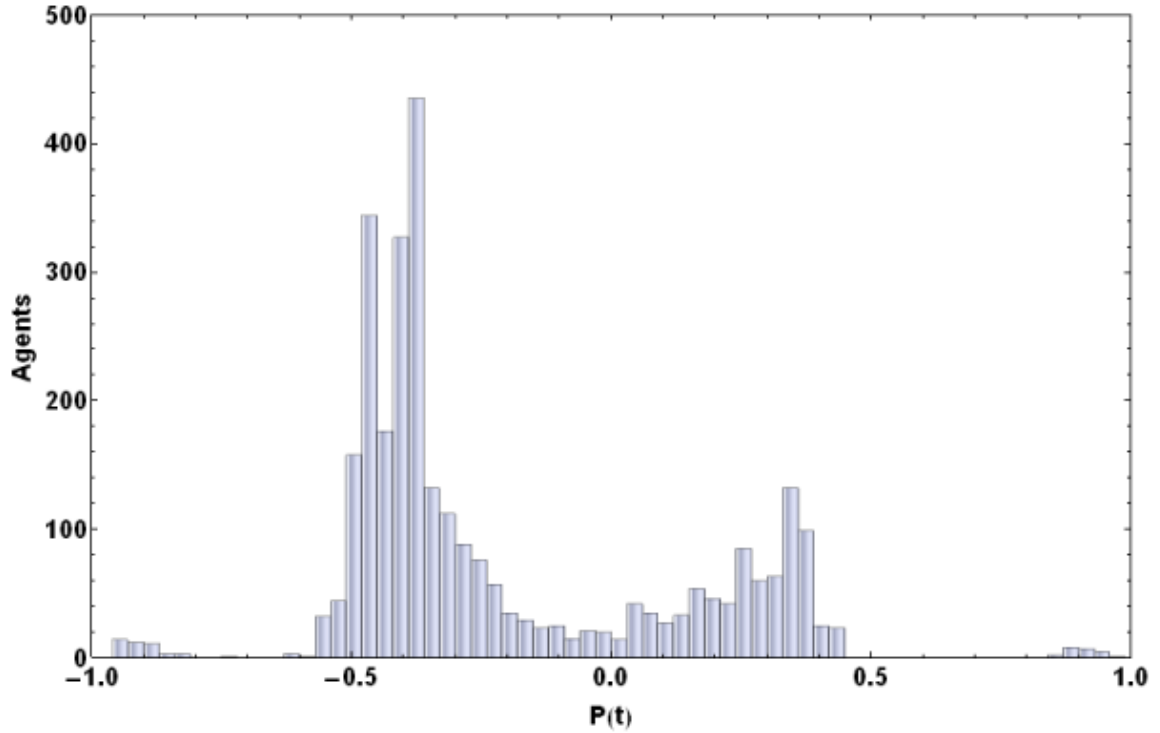


Figure 7: Multimodal, non-symmetric opinion distributions transiently appear as a result of combined homophily and attitude strength. The histogram of private opinions for this realization reveals a majority and a minority moderate party, a population of undecided centrists, and two small extremist parties.

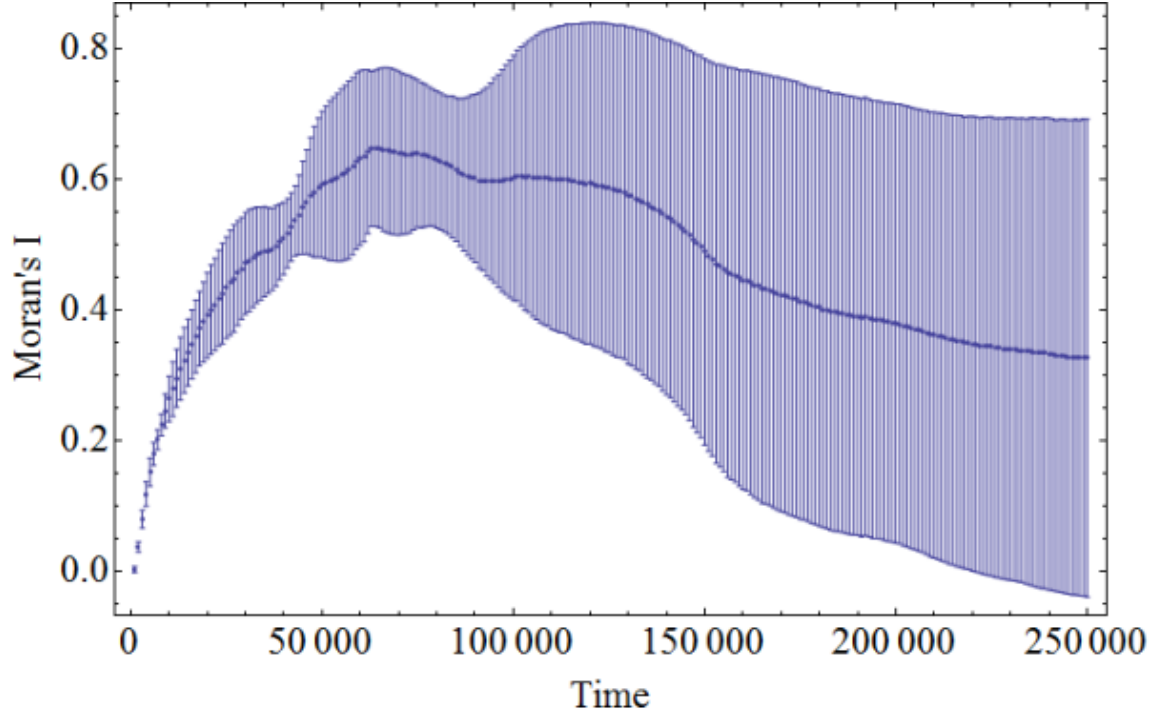


Figure 8: The spatial autocorrelation of agents opinions increases rapidly as neighborhoods coalesce and grow in size, then gradually declines as these neighborhoods mix on the borders and compete for territory. The large error (showing the standard deviation across 100 stochastic realizations) reflects the diversity of potential outcomes, which range from complete takeover of one opinion group ($M_I = 0.8$) to sustained competition between two geographically-consolidated groups ($M_I = 0.8 - 0.5$) to a spatial mixing of polarized extremists ($M_I = 0.2$).

neighborhoods to grow, shrink, or migrate. These results reiterate that extremists instigate neighborhood polarization but moderates are largely responsible for recruiting susceptible centrists and expanding the party's territory.

Finally, the reported dynamics may eventuate in qualitatively different outcomes, both across stochastic realizations with identical conditions and through variation of parameters. Under the conditions which produced Figures 6 – 9, society may (a) bifurcate into two moderate parties which constantly compete for the “votes” of a small centrist party, (b) reach a majority consensus while confining the minority to the geographic periphery, where they form a resilient island of self-reinforcing dissent, or (c) be completely overtaken by an uncompromising extremist ideology. Although each of these configurations may appear stable for long periods of time, the random aspects of dialogues and attitude change may still upset these equilibria and start society down a new trajectory. The diversity of outcomes across realizations reflects the randomness and path dependence of the simulations “political histories”.

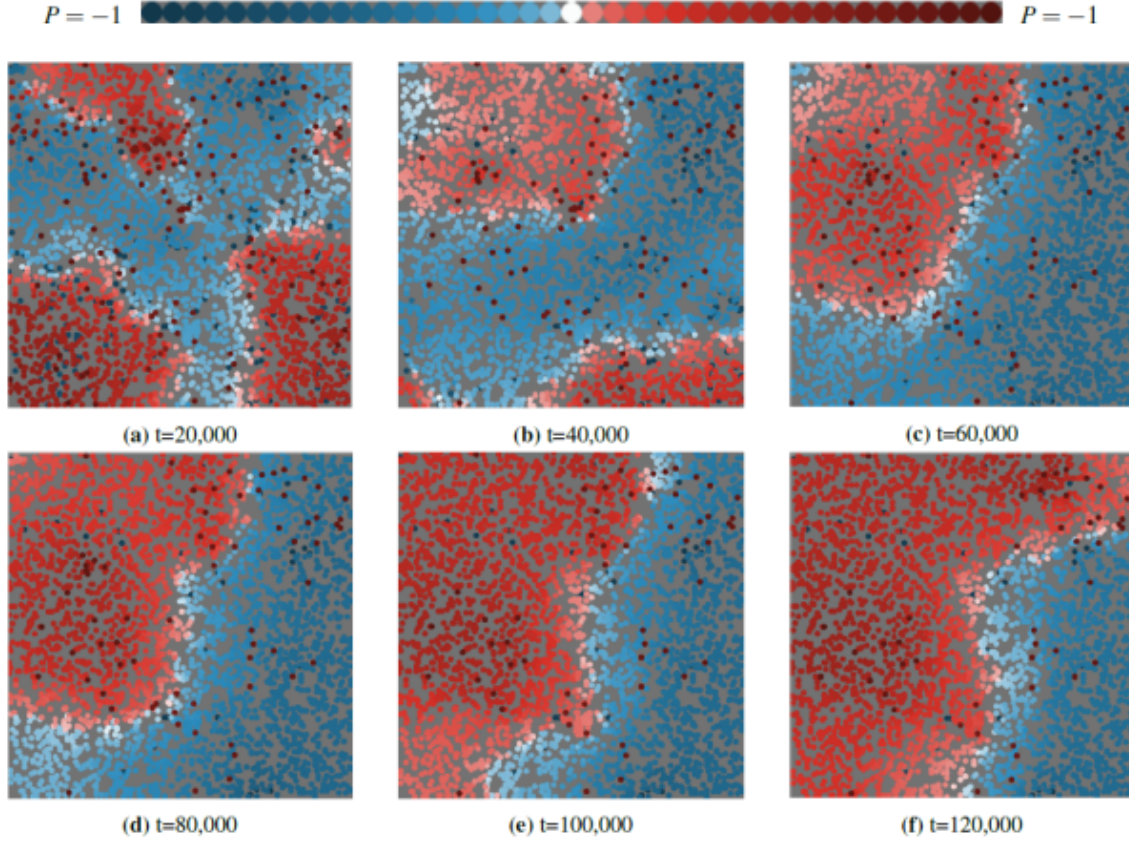


Figure 9: Neighborhoods of homogeneous opinions form by $t = 20,000$, but over the next 40,000 dialogues, some break apart, move towards centrism, or switch allegiance entirely. Around $t = 60,000$, two large neighborhoods coalesce in the NW and SE, and over the next 60,000 dialogues, these regions compete for the undecided centrists on their border. In the reported run, the blue region initially has a territorial and numerical advantage, but red influence makes a comeback and eventually out-competes the blue population. The political norms in some geographic regions, such as the SW corner, oscillate multiple times. In other runs, two competing neighborhoods may “chase” each other indefinitely around the map.

Conformity Conceals Attitude Diversity

As a final extension to the model, I introduce *preference falsification*, a term coined by Kuran (Kuran 1997) to describe individuals purposeful misrepresentation of their private beliefs in public settings. Conformity to perceived norms encourages preference falsification when an individuals desire to build and maintain meaningful social relationships outweighs his desire to remain consistent with a socially-undesirable personal belief (Cialdini and Goldstein 2004; Wood 2000). Where homophily represented external attributes of a stimulus and attitude strength represented the internal attributes of an agent, conformity represents how the contextual factors of a dialogue may affect agents reevaluation and expression of their political attitudes. This experiment investigates the effects of conformity/falsification by differentiating between opinions that agents *express in dialogues* and the attitudes they *privately hold*, a distinction that is critical in understanding many socio-political dynamics but has yet to be recognized in models of attitude change. I show that conformity-induced preference falsification can hide undercurrents of extremism in society in a manner that significantly affects the dynamics of societal attitude change.

An agent's *expressed opinion*, $E_i(t)$, is a function of his private belief, the social setting, and his desire to gain social approval. Specifically, his expressed opinion lies between his private beliefs and the political norm of his social networks:

$$E_i(t) = P_i(t) + c_i * (\overline{E_n(t)} - P_i(t)) \quad (9)$$

where c_i is agent i 's social conformity and $\overline{E_n(t)}$ is the mean expressed opinion of agent i 's network. Equation 9 implies that as agent i 's conformity rises $c_i \rightarrow 1$, his expressed opinion will move away from his true belief and towards the mean opinion of his network, $E_i(t) \rightarrow \overline{E_n(t)}$. Conformity c_i thus plays a role analogous to weight w_{ij} in Equations 5 and 7, but affects expressed rather than private opinions. When determining weight and impact, Equations 2, 5, and 7, agents now observe $E_j(t)$ rather than $P_j(t)$.

What factors pressure an individual to conform to social norms? Studies have shown that individuals with low self-esteem exhibit higher rates of social conformity (Gergen and Bauer 1967). Self-esteem, in turn, is closely related to whether individuals are accepted into valued reference groups (Turner et al. 1987). Individuals often assess the validity of their beliefs by comparing them to group norms, and suffer reductions in self-esteem if they learn the group holds a position counter to their own (Pool et al. 1998). Combining these ideas, I propose that individuals are more likely to conform to political norms when their private beliefs enjoy little support from members of their social network:

$$c_i = s_c * |P_i(t) - \overline{E_n(t)}| \quad (10)$$

where s_c , the salience of conformity, is a parameter governing agents' *reliance on social support*, and $\overline{E_N(t)}$ is the mean expressed opinion of agent i 's network.

I begin by introducing moderate social reliance ($s_c = 0.5$) into a run where opinions normally bifurcate into two moderate parties, Figure 1b. As before, opinions initially diverge and spatially mix, leading to high attitude heterogeneity within networks. High heterogeneity implies large disagreement between most agents private beliefs and network norms, and agents consequently experience low social self esteem. In an effort to “gain acceptance” by appealing to these norms, agents express conformist views center on $E_i(t) = 0$, the attitude position that best appeals to a heterogeneous audience.

With the majority of the population expressing centrist opinions, agents now perceive a strong centrist norm even though many agents still privately hold non-centrist beliefs. Figure 10 compares the number of agents who privately hold extremist vs. moderate beliefs to the number of agents who express extremist vs. moderate attitudes in dialogues. Although the number of agents who privately hold extreme beliefs is double that of moderates, at least ten times more moderate political expressions occur in dialogues between agents. Conformity thus *obscures undercurrents of extremism within society*, both to outside observers and to members of the population⁶. Figure 11 shows that while neighborhoods may verbally reach a consensus on a political issue, quiescent minorities can still be found mixed within each neighborhood, reiterating how unanimity of expressed opinions belies persistent diversity of privately held opinions.

Finally, preference falsification instigates a *private-public feedback* loop, in which perceived centrist norms inspire adoption of centrist private beliefs, and centrist private beliefs reinforce centrist public norms. The result is the “self-annihilation” of preference falsification⁷ as private and expressed opinions rapidly converge to $P(t) = E(t) = 0$, a markedly different outcome than the bifurcation of opinions into moderate parties obtained in the no-conformity case.

DISCUSSION

This study suggests three conclusions regarding the relationship between individual and societal attitude change. Firstly, *non-monotonic dynamics of attitude change are driven by localized interactions between closed-minded and persuadable individuals*. Although homophily and limited tolerance for dissenting political opinions are important principles governing attitude change, they are insufficient to produce non-linear behavior, opinion stability, and spatial clustering. After reproducing the classic relationship between tolerance and opinion convergence/divergence,

⁶These non-representative public norms are related to the psychological concept of pluralistic ignorance, where the majority of individuals privately reject a norm but nonetheless follow it because they assume that others accept it (Krech and Crutchfield 1948).

⁷As with the assignment of inter-agent weight in Equations 5 and 7, Equation 10 is a conceptually useful approximation that has been empirically shown to hold under some circumstances. Deepening it to include additional factors might allow, for example, agents who derive satisfaction from distinctiveness and maintained high self esteem despite receiving no support from their social networks.

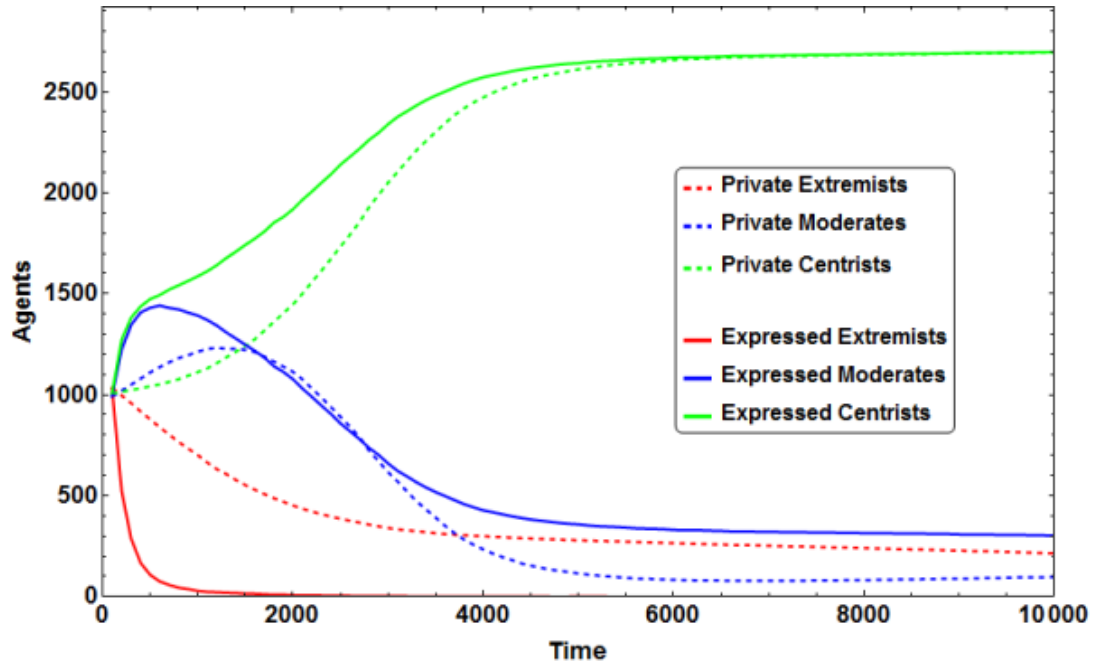


Figure 10: Moderate political statements that are expressed in political dialogues outnumber extremist statements by more than 10 to 1, as can be seen by comparing the blue vs. red solid lines after $t = 4000$. This statistic totally misrepresents the ratio of private opinion, blue vs. red dotted lines: nearly twice as many agents privately hold extreme beliefs as moderate beliefs. An observer judging the distribution of political preferences based solely on expressed opinions would significantly underestimate the extremist fraction within the population.

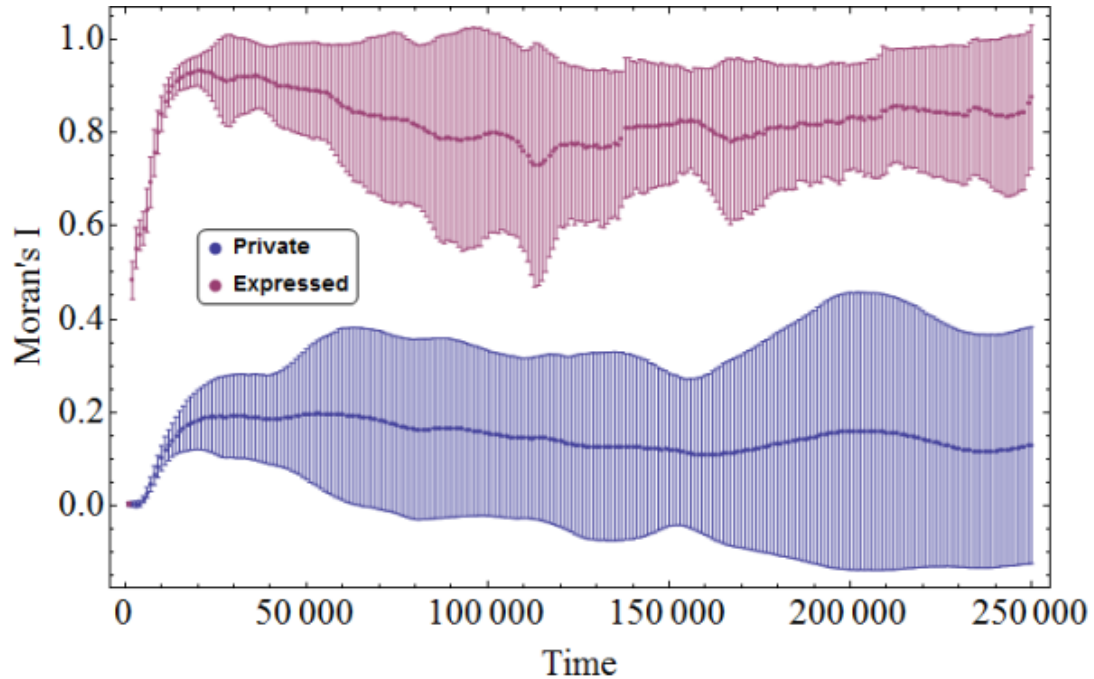


Figure 11: Conformity, motivated by agents' strong reliance on social support, produces neighborhoods of homogeneous expressed opinions, as shown by the high value of Moran's I in red. However, agents exist within these neighborhoods who privately disagree with the social norms but are too fearful of rejection to voice their dissent. Consequently, private opinions are more spatially mixed than expressed opinions, as shown by the lower value of Moran's I in blue.

the model showed how attitude rejector agents became more extreme with exposure to dissimilar opinions (attitude polarization) and persisted in their strong beliefs despite being surrounded by a centrist majority. Given enough time, these extremists polarized society without moderating their own beliefs, leading to an overall dynamic of centrist convergence followed by extremist polarization. When agents with strong beliefs underwent less attitude change than neutral agents, society self-organized into stable extremist and centrist parties. Combining homophily and attitude strength generated non-monotonic, dynamic interactions between extremist, moderate, and centrist parties, eventuating in a variety of outcomes including multimodality, non-symmetrical stability, and takeover. Finally, social conformity motivated by a lack of self-esteem created neighborhoods of preference falsification that fluctuated with the solidity of social norms and drove feedback between publicly expressed and privately held opinions. To my knowledge, this is the first generative model to produce both non-monotonic attitude dynamics and a persistent diversity of opinions across society, and therefore represents a significant advance in the scientific study of opinion formation and change.

Secondly, *neighborhood polarization is initiated by extremists and propagates through susceptible individuals*. This trend was observed in three experiments. In the first, attitude-rejector agents formed extremist neighborhoods and convinced susceptible, high-tolerance centrists within their networks to adopt moderate attitudes. These moderates went on to radicalize other proximate centrists, creating a wave of extremism that built enough momentum that it became self-sustaining, continuing even when the extremists were artificially removed. In the second experiment, attitude-strength encouraged extremists to stubbornly reject all influence and increased the susceptibility of centrist agents to attitude change. When combined with homophilous dynamics, extremist neighborhoods again formed and polarized nearby susceptible centrists. Finally, in neighborhoods where moderate expressions from one extremist party outweighed expressions from the minority group, conformity drove local consensus, forming clusters of homogeneous, moderate expressed opinions. A self-reinforcing cycle emerged in which attitude expressions became more extreme, extremists expressed more radical positions, and conformity to new radical norms inspired further attitude change. These results emphasize the importance of understanding opinion change from both a opinion-distribution perspective and a geographic/network perspective.

Lastly, *preference falsification obscures attitude diversity and consequently alters dynamics of attitude change*. When agents conform to social norms, undercurrents of extremism are not only hidden to outside observers – they are unseen by agents as well. Under conditions that otherwise resulted in bifurcation of opinions into moderate parties, conformity masked societal diversity and instead caused rapid convergence to a centrist opinion, but also isolated a previously-incorporated extremist fringe. Strong conformity induced feedback between expressed opinions (public political norms) and private opinions (personal political beliefs), producing volatile clusters of expressed opinions that oscillated between extremism, moderatism, and centrism⁸. Integrating preference falsification into models of attitude change is essential to discerning the origins of “tipping points” in public

opinion, such as the legalization of marijuana and gay marriage in the USA or the unanticipated nature of political revolutions (Kuran 1989).

Applications to Conflict Resolution

Consider a society where strong disagreement on a political issue is a source of conflict among extremist factions. What programs could individuals, neighborhoods, or governments implement to depolarize radical groups and manage this conflict? My findings suggest why several simple programs may be ineffective. First, starting dialogues with extremists and presenting them with reasoned counterarguments might be persuasive to purely rational individuals, but emotionally charged extremists often assimilate evidence in a biased manner. If counterarguments are presented inappropriately, extremists may undergo attitude polarization and adopt more extreme attitudes. Second, breaking up radical groups and integrating extremists into politically diverse neighborhoods can also backfire. Depending on the extent of similarity between individuals, their political tolerance, and their open-mindedness, extremists may polarize a moderate neighborhood rather than being depolarized by it. Third, segregating radicals may reduce conflict in the short term, but isolated communities of extremists can grow more radical over time. A dense pocket of extremists, even if small and confined to the periphery, can instigate conflict on its borders and polarize surrounding neighborhoods. Finally, quelling extremism by legally or socially sanctioning the expression of radical attitudes may prevent conflict from surfacing in public spaces, but precludes open dialogues that might otherwise lead to depolarization. Silencing dissent also conceals undercurrents of extremism, which may accumulate and violently resurface in response to an external shock.

Validation

The following experiment quantifies how individuals modify their opinions on a political issue when presented with diverse opinion statements. A sample of human subjects, heterogeneous in their political preferences, take a pre-test survey that quantifies their opinion on a political issue⁹ using a numerical scale and prepare short statements expressing their perspective on the issue. These data are used to initialize a population of agents in the model with heterogeneous beliefs, scaled to $(1, 1)$. Participants are then given a series of each others prepared statements to read and consider. The attitudes expressed in these statements are quantified by a second experimental group and are used to calibrate the E_{ij} s communicated between agents during dialogues. After reading and reflecting on these

⁸Private-public feedback also caused preference falsification to “self-annihilate”, a result that is perhaps most apparent across multiple generations. When strong pressures to conform lead parents to hide their private beliefs from the public, they may nonetheless resist privately adopting social norms due to personal history, inherently high self-esteem, or secret networks of non-conformists. However, their children will be raised in a society with strong social norms and high pressures to conform: because they are rarely exposed to non-conformist viewpoints and lack the life experiences that would inspire self-esteem and resilience to social rejection, these children will privately adopt the political norms of society. When the older generation dies off, political dissent disappears.

⁹The issue under investigation should be contentious, in the sense that society contains a diverse set of private and public opinions, and subjective; e.g. abortion or capital punishment.

attitude statements, participants retake the opinion-identification survey and modify their initial statements to reflect changes in their opinions ¹⁰. Finally, the model is run with the appropriate $P_i(0)$ and $\sum_j^N E_j$ for each agent, and agents final private opinions $P_i(t)$ are compared with experimental participants final opinions. This experiment tests whether the model successfully predicts the extent of individuals attitude change given information on initial opinions and the dialogues that occur between people.

Extensions

The generative methodology employed in this study can be summarized as follows: (1) begin with a simple, theoretically and empirically grounded formalism for studying some phenomena of interest; (2) introduce one new component to the model, explain the rationale behind the addition, and provide alternatives and critiques when appropriate; (3) use computational experiments to identify the resulting behaviors, averaged over stochastic realizations with heterogeneous agents; (4) dig deeper to determine the chain of causality responsible for the model result, corroborated using mathematical proofs or alternate assumptions; (5) relate the results back to theory, stylized facts, or empirical data; (6) if the addition increases the models explanatory power, keep it, and if not, discard it. This modeling approach could be extended by (a) simulating agent movement, the formation and dissolution of network connections, and modern communication technologies; (b) endowing agents with belief webs in which the relationships between political issues are modified by personal experiences; (c) developing agents social and propositional reasoning abilities ¹¹.

ACKNOWLEDGMENTS

I would like to thank my coworkers at the Johns Hopkins Center for Advanced Modeling in the Social, Behavioral, and Health Sciences for their guidance and feedback on this project. In particular, I thank Erez Hatna for our many discussions about the model’s structure, metrics, and results, Paul Smaldino, Michael Makowsky, and Brett Calcott for their ideas and editorial feedback, and Joshua Epstein for providing the opportunity to work in such a stimulating and supportive setting, as well as for the thought-provoking ideas he presents in *Agent Zero: Toward Neurocognitive Foundations for Social Science*.

¹⁰A variant of the experiment divides participants into small groups where they present their statements one-at-a-time. Before each participant makes his statement, he is given a chance to modify it based on the statements he has heard thus far. His prepared statement, his expressed statement, and his post-dialogue statement are all recorded, allowing the quantification of preference falsification.

¹¹The heuristics of homophily, attitude strength, and conformity are approximations for underlying cognitive processes that constructively and destructively interfere in unpredictable ways. Gaining a deep understanding of how, when, and why these heuristics emerge requires building a generative model of human cognition. Neurocognitive agents must (a) form attitudes based on personal experience and second-hand evaluations (b) store these evaluations in a semantic, connectionist memory (c) selectively retrieve/reconstruct these attitudes when presented with the appropriate stimuli, and (d) use affective, social, and propositional reasoning to conduct higher-order political analyses. Building such a model is a tall order, but the applications to social science would be innumerable.

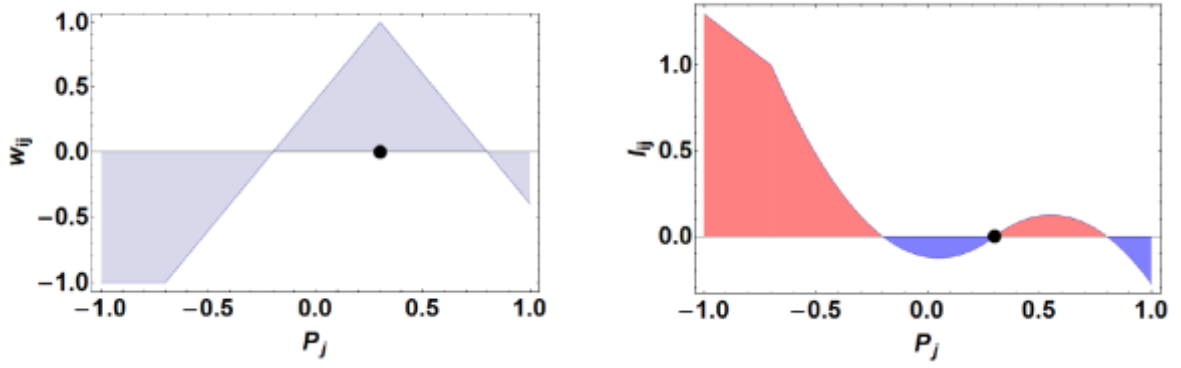


Figure 12: Attitude polarization results from the rejection of dissimilar attitudes. The x-axes represent the opinion that all agents j express in a dialogue, and the black dot shows the initial opinion of agent i . The weight (Equation 5 and impact (Equation 2 that agent i assigns to each agent j 's opinion are shown on the left- and right-y-axes, respectively.

Appendix A: Attitude Rejection

Figure 12 graphically demonstrates why rejector agents undergo attitude polarization when exposed to a uniform distribution of opinions. Suppose an attitude rejector agent with positive initial opinion, $P_i(0) = 0.3$, is exposed to a continuous uniform distribution of opinions P_j , shown on the x-axes. The rejector agent assigns to each P_j a weight w_{ij} , shown on the left, which is negative if it differs significantly from his own, $\Delta P_{ij} > \frac{1}{2}$ for $s_h = 2$. When w_{ij} is positive, the agent will shift his private opinion in the direction of the statement P_j , otherwise he will shift it in the opposite direction. The impact I_{ij} resulting from each P_j is shown on the right; integrating over all P_j gives the total directed impact experienced by the agent after exposure to the uniform opinion distribution. Regions of P_j which cause impact towards $P = 1$ are colored red and regions which cause impact towards $P = -1$ are colored blue. Clearly, red impact dominates, and the agent adopts a more positive opinion than he initially held, $P_i(t) \rightarrow 1$. In general, rejector agents who are exposed to a even-handed distribution of opinions will reject those farthest away from them, and this rejection will dominate their total impact¹², causing *attitude polarization* and the adoption of more radical views.

To show that attitude rejection is responsible for the observed neighborhood polarization, I alter the model to reproduce the spatial clustering observed in Jager's ABM of attitude change (Jager and Amblard 2005). Equation 5

¹²The plot on the right also shows the quadratic influence of $\Delta P_{ij}(t)$ on impact $I(t)$, which can be obtained by substituting Equation 5 into Equation 2; the same quadratic occurs between expressed opinion and $\overline{E_n(t)} - P_i(t)$. The strength of this dependence is greater than in other opinion dynamics models.

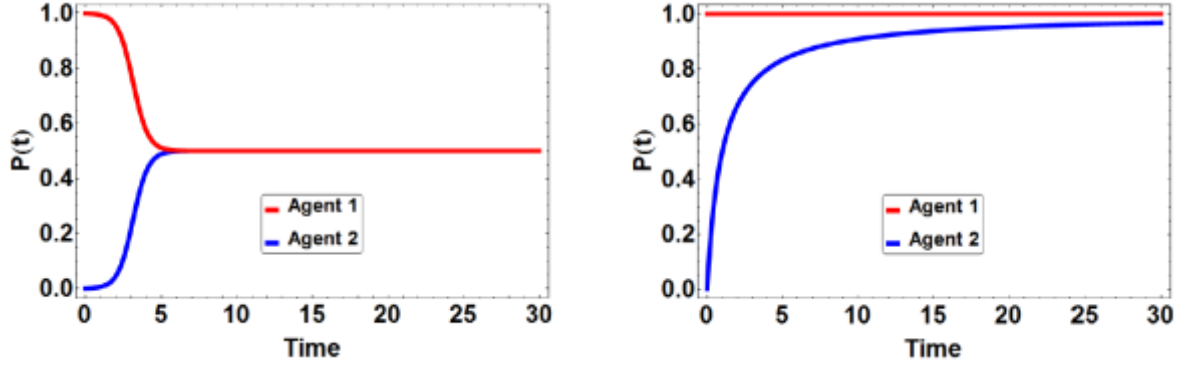


Figure 13: Convergence of opinions in a two agent system when agents employ homophilous weighting (left, Equation 5) vs. the attitude strength weighting (right, Equation 7). Homophily causes agents to converge symmetrically to a central opinion within $t = 5$, while attitude strength causes a decaying, non-symmetrical convergence towards the impact-resistant extreme agent.

becomes

$$w_{ij} = \begin{cases} 0.5 & \text{if } \Delta P_{ij}(t) < \alpha \\ 0 & \text{if } \alpha \leq \Delta P_{ij}(t) \leq \beta \\ -0.5 & \text{if } \Delta P_{ij}(t) > \beta \end{cases} \quad (11)$$

where α is the acceptance threshold and β is the rejection threshold. Negative w_{ij} indicates that agent- j 's opinion statement has a negative impact on agent- i , causing the latter to shift $P_i(t)$ away from $P_j(t)$. Each cell on a 50×50 lattice is occupied by an agent, whose social network is reduced to their Moore neighborhood. With $\beta > 2$, rejection is impossible, and varying α produces convergence or divergence with spatial mixing. With rejection, $\beta = 1.5$ and $\alpha = 1.0$, polarized neighborhoods emerge and grow in size until they dominate the population; Moran's I increases logarithmically from 0 to 0.8 ± 0.1 over $t = 300,000$ dialogues. As suspected, rejection is responsible for neighborhood polarization, even under different assumptions regarding social networks and homophily.

Figure 13 shows the opinion trajectories for two interacting agents employing Equations 5 and 7 respectively. As $P_{blue}(t) \rightarrow \frac{1}{s_a}$, $w_{ij} \rightarrow 0$, meaning that the rate at which a moderate's opinion converges to an extremist's opinion decays with ΔP_{ij} . In finite time $P_{blue}(t)$ will never reach $P_{red}(t)$. Similarly, as $P_{blue}(t) \rightarrow 0$, $w_{ij} \rightarrow 1$, implying that agents approaching centrism will become strongly susceptible to influence and likely be perturbed back towards moderatism.

Appendix B: Parameter Values

Name	Figure	Conditions	Result
Convergence	1a	$s_h = 1.0$	absolute centrist convergence
Divergence	1b	$s_h = 2.5$	absolute moderate bifurcation
Mixing	2	$s_h = 2.5$	spatial mixing at network level
Clustering 1	3	$s_h = 1.25, N_r = 15\%$	neighborhood polarization
Nonmonotonic 1	4	$s_h = 1.25, N_r = 15\%$	initial centrist convergence, long-term extremist takeover
Diversity	5	$s_a = 1.25$	persistent diversity of opinions
Nonmonotonic 2	6	$s_h = s_a = 0.75$ ($t = 100,000$)	non-monotonic attitude dynamics, many potential outcomes
Multimodal	7		multimodal, non-symmetric opinion distributions
Clustering 2	8		nonmonotonic spatial autocorrelation of opinions
Clustering 3	9	(seed = 5)	neighborhood polarization, growth, competition for centrists
Pluralistic Ignorance 1	10	$s_h = 2.0, s_c = 0.5$	convergence and hidden extremist undercurrents
Pluralistic Ignorance 2	11		public consensus with hidden private dissent
Attitude Polarization	12	$s_h = 2.0$	opinions grow stronger with exposure to dissimilarity
Two agent system	13a	$s_h = 1.0$	symmetrical convergence
	13b	$s_a = 1.0$	non-symmetrical convergence

Table 1: List of figures with corresponding parameter values and major findings

References

- Ajzen, I. (2001). Nature and operation of attitudes. *Annual review of psychology* 52(1), 27–58.
- Betz, A. L., J. J. Skowronski, and T. M. Ostrom (1996). Shared realities: Social influence and stimulus memory. *Social Cognition* 14(2), 113–140.
- Bohner, G. and N. Dickel (2011). Attitudes and attitude change. *Annual review of psychology* 62, 391–417.
- Cialdini, R. B. and N. J. Goldstein (2004). Social influence: Compliance and conformity. *Annu. Rev. Psychol.* 55, 591–621.
- Dandekar, P., A. Goel, and D. T. Lee (2013). Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences* 110(15), 5791–5796.
- Deffuant, G., F. Amblard, G. Weisbuch, and T. Faure (2002). How can extremism prevail? a study based on the relative agreement interaction model. *Journal of Artificial Societies and Social Simulation* 5(4).
- Deffuant, G., D. Neau, F. Amblard, and G. Weisbuch (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems* 3, 87–98.
- DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association* 69(345), 118–121.
- Gawronski, B. and G. V. Bodenhausen (2006). Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. *Psychological bulletin* 132(5), 692–731.
- Gergen, K. J. and R. A. Bauer (1967). Interactive effects of self-esteem and task difficulty on social conformity. *Journal of personality and social psychology* 6(1), 16–22.
- Hamill, L. and N. Gilbert (2010). Simulating large social networks in agent-based models: A social circle model. *Emergence: Complexity & Organization* 12(4), 78–94.
- Hegselmann, R. and U. Krause (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation* 5(3).
- Jager, W. and F. Amblard (2005). Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change. *Computational & Mathematical Organization Theory* 10(4), 295–303.
- Krech, D. and R. S. Crutchfield (1948). *Theory and problems of social psychology*. McGraw-Hill Book Company.

- Kuran, T. (1989). Sparks and prairie fires: A theory of unanticipated political revolution. *Public Choice* 61(1), 41–74.
- Kuran, T. (1997). *Private truths, public lies: The social consequences of preference falsification*. Harvard University Press.
- Latane, B. (1981). The psychology of social impact. *American psychologist* 36(4), 343–356.
- Lord, C. G., L. Ross, and M. R. Lepper (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology* 37(11), 2098–2109.
- McPherson, M., L. Smith-Lovin, and J. M. Cook (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology* 27, 415–444.
- Miller, A. G., J. W. McHoskey, C. M. Bane, and T. G. Dowd (1993). The attitude polarization phenomenon: Role of response measure, attitude extremity, and behavioral consequences of reported attitude change. *Journal of Personality and Social Psychology* 64(4), 561–574.
- Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika* 37(1-2), 17–23.
- Morris, A. and S. Staggenborg (2004). Leadership in social movements. *The Blackwell companion to social movements*, 171–196.
- Petty, R. E., D. T. Wegener, and L. R. Fabrigar (1997). Attitudes and attitude change. *Annual review of psychology* 48(1), 609–647.
- Pool, G. J., W. Wood, and K. Leck (1998). The self-esteem motive in social influence: agreement with valued majorities and disagreement with derogated minorities. *Journal of personality and social psychology* 75(4), 967–975.
- Salzarulo, L. (2006). A continuous opinion dynamics model based on the principle of meta-contrast. *Journal of Artificial Societies & Social Simulation* 9(1).
- Taber, C. S. and M. Lodge (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science* 50(3), 755–769.
- Turner, J. C., M. A. Hogg, P. J. Oakes, S. D. Reicher, and M. S. Wetherell (1987). *Rediscovering the social group: A self-categorization theory*. Basil Blackwell.
- Visser, P. S. and J. A. Krosnick (1998). Development of attitude strength over the life cycle: surge and decline. *Journal of personality and social psychology* 75(6), 1389–1410.

- Wilensky, U. (1999). Netlogo. *Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.*
- Wood, W. (2000). Attitude change: Persuasion and social influence. *Annual review of psychology* 51(1), 539–570.
- zu Erbach-Schoenberg, E., S. Bullock, and S. Brailsford (2013). A model of spatially constrained social network dynamics. *Social Science Computer Review* XX(X), 1–20.